

The Arabidopsis Unannotated Secreted Peptide Database, a Resource for Plant Peptidomics^[W]

Kevin A. Lease* and John C. Walker

Division of Biological Sciences, University of Missouri, Columbia, Missouri 65211

In the era of genomics, if a gene is not annotated, it is not investigated. Due to their small size, genes encoding peptides are often missed in genome annotations. Secreted peptides are important regulators of plant growth, development, and physiology. Identification of additional peptide signals by sequence homology searches has had limited success due to sequence heterogeneity. A bioinformatics approach was taken to find unannotated Arabidopsis (*Arabidopsis thaliana*) peptides. Arabidopsis chromosome sequences were searched for all open reading frames (ORFs) encoding peptides and small proteins between 25 and 250 amino acids in length. The translated ORFs were then sequentially queried for the presence of an amino-terminal cleavable signal peptide, the absence of transmembrane domains, and the absence of endoplasmic reticulum luminal retention sequences. Next, the ORFs were filtered against the The Arabidopsis Information Resource 6.0 annotated Arabidopsis genes to remove those ORFs overlapping known genes. The remaining 33,809 ORFs were placed in a relational database to which additional annotation data were deposited. Genome-wide tiling array data were compared with the coordinates of the ORFs, supporting the possibility that many of the ORFs may be expressed. In addition, clustering and sequence similarity analyses revealed that many of the putative peptides are in gene families and/or appear to be present in the rice (*Oryza sativa*) genome. A subset of the ORFs was evaluated by reverse transcription-PCR and, for one-fifth of those, expression was detected. These results support the idea that the number and diversity of plant peptides is broader than currently assumed. The peptides identified and their annotation data may be viewed or downloaded through a searchable Web interface at peptidome.missouri.edu.

Peptide signals control diverse aspects of plant physiology, growth, and development. For example, there are demonstrated roles for peptide signals in defense responses (McGurl et al., 1992; Huffaker et al., 2006), maintenance of stem cell identity in the shoot apical meristem (Fletcher et al., 1999; Kondo et al., 2006), self-incompatibility in crucifer species (Schopfer et al., 1999; Takayama et al., 2000), and cell proliferation and differentiation (Matsubayashi and Sakagami, 1996; Matsubayashi et al., 1999; Ito et al., 2006). These signaling peptides act through cell surface receptors to elicit physiological responses. In addition, there are also peptides that act as direct effectors, such as antimicrobial peptides involved in innate immunity or phytochelators that limit abiotic stress due to toxic metals (Garcia-Olmedo et al., 1998; Cobbett and Goldsbrough, 2002).

Elucidating the roles of additional plant peptides is essential to understanding the intercellular communication underlying plant biology. A bottleneck in the process of establishing the functions of plant peptides has been identifying the genes that encode peptides. This is because many peptides are encoded by small genes, which tend to be missed in genome annotations.

One piece of evidence suggesting that peptide-encoding small genes are underannotated in Arabidopsis (*Arabidopsis thaliana*) includes several examples of large gene families of plant peptides missed in the genome annotation. In such cases, the identification of a founding peptide allowed for additional unannotated family members to be recognized by sequence homology searches (Vanoosthuysen et al., 2001; Olsen et al., 2002; Wen et al., 2004).

What are some possible reasons for the decreased frequency of annotating small genes? Genes are discovered by a combination of gene-finding computer programs that analyze genomic DNA sequences, cDNA sequencing, and by characterization of mutants. In computational gene identification, to minimize incorrectly predicting random small open reading frames (ORFs) to be genes, it is common practice to disregard ORFs below a certain size without empirical evidence of expression (i.e. an expressed sequence tag or complementary DNA). This practice seeks to optimize the signal-to-noise ratio in gene finding, reflecting that the probability of a spurious ORF in the genome occurring by chance increases as the size of the ORF decreases. Bias also is a result of the methods used to construct cDNA libraries, which typically utilize a molecular size selection step to eliminate cDNAs smaller than 400 to 500 bp. Genetic screens can potentially uncover small genes; however, the small gene size would work against its discovery by reducing the likelihood of a mutagenic event occurring within that gene. In combination, these factors contribute to the underreporting of small peptide-encoding genes in genome annotations.

Based on the aforementioned reasoning, we hypothesized that additional small genes encoding peptides

* Corresponding author; e-mail leasek@missouri.edu; fax 573-884-9676.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Kevin A. Lease (leasek@missouri.edu).

^[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.106.086041

remain to be discovered. We used bioinformatic approaches to both provide additional evidence that small peptides are underannotated and to identify candidate peptide-encoding genes that are currently unannotated.

RESULTS

Protein Length Frequencies Suggest Small Proteins May Be Underannotated

One expectation, if small genes encoding peptides are underannotated in Arabidopsis, is that there would be a skew in the protein length frequency distribution. To evaluate whether there is a disparity in the number of annotated proteins of different lengths, we plotted the frequency of The Arabidopsis Information Resource (TAIR) 6.0 annotated Arabidopsis proteins as a function of length, after placing them in bins 50 amino acids wide (Fig. 1). This graph indicates a substantial drop off in the number of small proteins less than 101 amino acids, especially below 51 amino acids. It is even more striking when one considers that the number of theoretically possible ORFs increases exponentially as the size of the ORF allowed decreases (Basrai et al., 1997). The observed distribution may either reflect a biological basis (i.e. there really are relatively fewer genes encoding small proteins in the Arabidopsis genome) or it may indicate an underannotation of small protein-coding genes.

Bioinformatic Approach to Identify Unannotated Secreted Peptides

Because our interests are in intercellular signaling, we focused upon small genes encoding peptides that are likely to be secreted. We wrote Perl scripts to automate the process of unannotated Arabidopsis peptide identification and sequential filtering, the results of which are summarized in Table I. First, we screened all five Arabidopsis chromosomes on both strands for all ORFs encoding proteins 25 to 250 amino acids in length. The rationale for screening for ORFs was that many known plant peptides believed to function in intercellular signaling are encoded by single-exon genes. For example, all of the CLV3/ESR-related (CLE) gene family (except CLV3 and CLE40; Cock and McCormick, 2001; Sharma et al., 2003), the rapid alkalization factor-like (RALFL) gene family (Olsen et al., 2002), the inflorescence deficient in abscission (IDA) family (Butenko et al., 2003), the DEVIL family (Wen et al., 2004), the LCR family (Vanoosthuyse et al., 2001), and the SCRL family (Vanoosthuyse et al., 2001) are single-exon genes. Also, it is worth noting that single-exon genes are the most frequently observed gene structure in Arabidopsis (Alexandrov et al., 2006). The lower size limit is based on the idea that the average signal peptide is 22 amino acids long (Bendtsen et al., 2004b) and the knowledge that there are bioactive plant peptides as short as five amino acids long (Matsubayashi and Sakagami, 1996). The upper size limit was set at 250

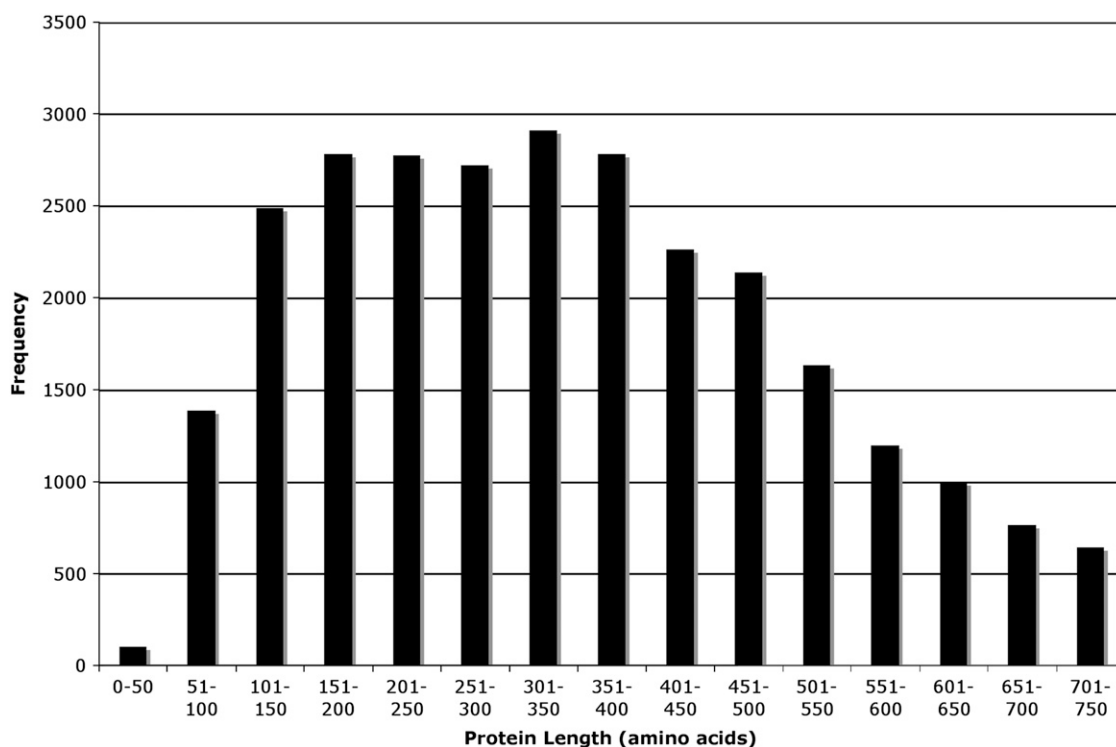


Figure 1. Asymmetric distribution observed in histogram of annotated Arabidopsis protein lengths. The lengths of the 30,690 proteins annotated in TAIR 6.0 were evaluated using a Perl script and placed in bins 50 amino acids wide. Bins containing proteins up to 750 amino acids were plotted.

Table 1. Bioinformatic identification and filtering of *Arabidopsis* peptide-encoding ORFs

Arabidopsis Chromosome	No. Identified ORFs 25–250 Amino Acids in Length	No. ORFs Kept after Each Sequential Filter				
		Keeping the Larger ORFs if Overlapping In-Frame ORFs	Presence of SignalP 3NN Predicted Cleavable Amino-Terminal Signal Peptide	Absence of TMHMM2 Predicted Membrane-Spanning Domains	Lack of KDEL or HDEL at C Terminus	Outside Gene Span of TAIR 6.0 Annotated Genes
1	153,649	120,233	21,253	19,811	19,810	8,214
2	99,714	78,754	13,935	13,022	13,022	6,289
3	121,487	94,968	16,324	15,327	15,327	6,812
4	94,585	74,185	12,713	11,904	11,904	5,301
5	136,850	107,141	18,472	17,206	17,206	7,193
Total	606,285	475,281	82,697	77,270	77,269	33,809

amino acids, based on the tomato (*Lycopersicon esculentum*) systemin precursor, the largest known signaling proprotein at 200 amino acids long, to allow for some tolerance. Screening the *Arabidopsis* genome with these constraints resulted in identification of 606,285 ORFs.

Next, we removed smaller ORFs if multiple overlapping in-frame ORFs were recovered, resulting in 475,281 ORFs retained. Predicted peptides and small proteins were screened for the presence of predicted amino-terminal signal peptides that would direct them to the secretory pathway, using the neural network version of SignalP 3.0 (Bendtsen et al., 2004b). This reduced the number of ORFs to 82,697.

We then screened for the absence of transmembrane helices that could indicate the protein resides in the plasma membrane or an endomembrane of the secretory pathway, using TMHMM 2.0 (Krogh et al., 2001), keeping 77,270 ORFs. We also eliminated possible endoplasmic reticulum lumen-resident protein by removing proteins having KDEL or HDEL motifs at their C termini (one sequence removed). Last, we screened against ORFs that overlapped the coordinates of genes annotated in TAIR 6.0 annotated genome release, resulting in final identification of 33,809 putative ORFs.

Bioinformatic Approaches to Validate Putative ORFs

We focused upon three properties of the predicted ORFs that could be assessed using bioinformatics to address ORF predictions. The first expectation of a functional gene is that it is transcribed. We used the published data generated in genome-wide tiling hybridization experiments (Yamada et al., 2003) to provide evidence of ORF expression. Specifically, we tested whether hybridization intensities were at least twice the median signal value of the chip. By this rationale, from the analysis of the root, leaf, flower, and suspension cell datasets, we found evidence for the expression of 10,247 putative ORFs.

Because most known peptide-encoding genes belong to gene families, a second predictive measure used was that putative peptides that are encoded by gene families are more likely to be valid. We used BLASTCLUST (Altschul et al., 1990) to perform single-linkage clustering of the preproteins (the protein with its signal peptide). This analysis revealed that 3,324 of the peptides belong to 974 clusters of two or more peptides.

Third, we reasoned that putative ORFs are more likely to represent genes if an ortholog can be found in the rice (*Oryza sativa*) genome. Using tBLASTn with default parameters, 15,975 putative ORFs were found to have a match to the rice genome. This threshold was chosen to allow for sensitivity, given the short length of peptides.

An evaluation of the degree of overlap of these three types of annotation data revealed that 1,044 predicted ORFs were supported by all three, and a further 6,042 predicted ORFs were supported by any two of the three (Fig. 2).

Reverse Transcription-PCR Shows a Subset of ORFs Tested Is Expressed

We chose 25 ORFs at random from the collection of ORFs having supporting data from all three classes of annotation data for testing by reverse transcription (RT)-PCR. Given that many genes exhibit differential expression, we separately isolated RNA from 2-week-old roots, 3-week-old rosette leaves, central aerial portions of 2-week-old plants (to get vegetative meristem tissue), flowers, and siliques. The lack of introns in the ORFs precluded utilizing primers that flank

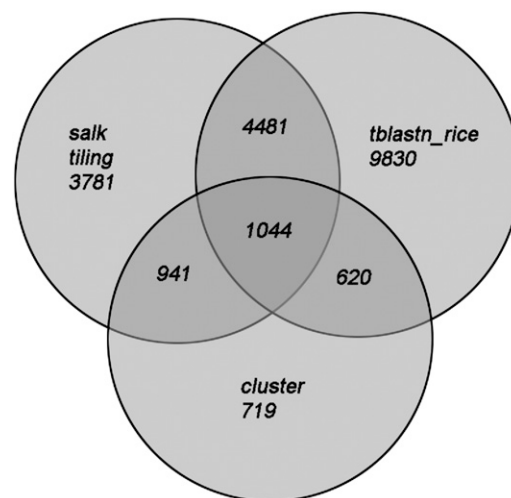


Figure 2. Venn diagram displaying the degree of overlap of annotation data supporting predicted peptides.

an intron as a safeguard against amplifying genomic DNA contamination. Two alternative measures were taken to both avoid and control for the possibility of genomic DNA contamination amplification. First, we treated the RNA with DNase before synthesizing the cDNA to remove any traces of genomic DNA. Second, we set up a mock cDNA synthesis reaction for each RNA in which no reverse transcriptase was added. PCR reactions using aliquots of the mock cDNA reactions were used as a template to demonstrate that PCR products were reverse transcriptase dependent. From these experiments, we found evidence for the expression of five of 25 tested ORFs (Fig. 3). The expected PCR products could be generated with all 25 primer pairs using genomic DNA as a template, indicating that the failure to detect expression with the cDNA template was not due to PCR conditions, primer design, or synthesis (data not shown). The ORFs and primers used that did not give evidence of expression are presented in Supplemental Table S1.

Searchable Web Interface for Dissemination of the Arabidopsis Peptidome Dataset

A MySQL relational database with information about the 33,809 ORFs was generated. Peptides were named with the prefix *ath_mu* to denote Arabidopsis predictions and the University of Missouri. Following this prefix, the chromosome, the number of the prediction, and the strand are contained within the name (e.g. *ath_mu_ch1_20337top*). A Web interface was designed that allows for querying the database for subsets of the

predicted peptides based on a variety of selection criteria (Fig. 4). The Web site address is <http://peptidome.missouri.edu>. Data may either be retrieved in tabular format in the Web browser window or may be downloaded to the desktop in a tab-delimited format, which can be opened in Microsoft Excel.

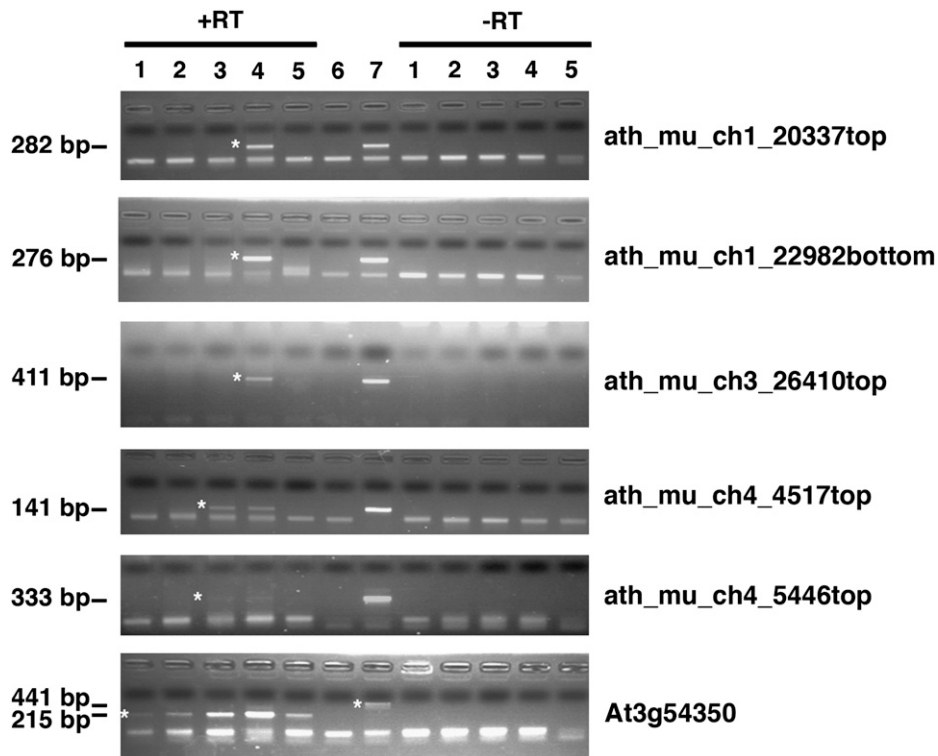
New Members of the RALFL Family Identified

We used sequence homology searches to test whether any peptides similar to known plant peptides were in the Arabidopsis Unannotated Secreted Peptide Database. We identified members of several known plant peptide families. For example, 12 putative RALFL (Pearce et al., 2001b; Olsen et al., 2002) genes were found. Among these 12, 10 were previously reported (Olsen et al., 2002) but are not included in the most recent version of the Arabidopsis genome annotation. A list of these ORFs is given in Supplemental Table S2. An additional two RALFL genes (*ath_mu_ch1_20831bottom*, *ath_mu_ch1_21704top*) that have not been previously reported were identified, expanding the number of RALFL genes in Arabidopsis from 34 to 36 (Fig. 5).

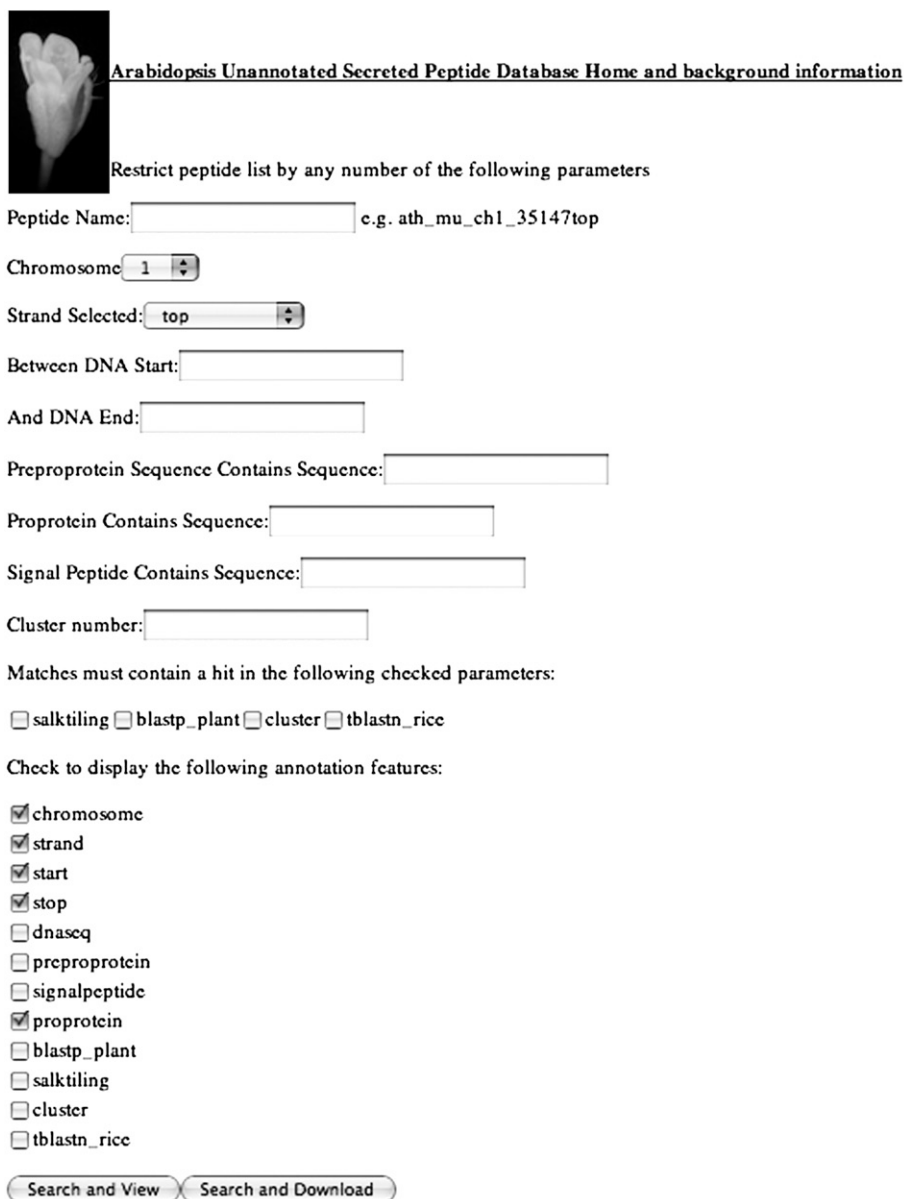
In addition, a recent addition to the CLE family, CLE45 (Strabala et al., 2006), was found in the database as *ath_mu_ch1_65447top*. Furthermore, *ath_mu_ch1_59545bottom* exhibits similarity to the CLE domain of CLE26 and may represent an unreported CLE.

Although they were described by Butenko et al. (2003), four inflorescent deficient in abscission-like (IDL) members have not yet been added to the official annotation of Arabidopsis genes and were identified

Figure 3. RT-PCR results showing expression of unannotated ORFs. Lane 1, Root; lane 2, leaf; lane 3, vegetative meristem; lane 4, flower; lane 5, fruit; lane 6, water negative control; lane 7, genomic DNA positive control. Asterisks indicate PCR products. *At3g54350* gene-specific primers were used as a positive control for first-strand cDNA synthesis and span two introns giving different size products for genomic and cDNA templates. Primer dimers are seen below PCR products.



Arabidopsis Unannotated Secreted Peptide Database Search



[Arabidopsis Unannotated Secreted Peptide Database Home and background information](#)

Restrict peptide list by any number of the following parameters

Peptide Name: c.g. ath_mu_ch1_35147top

Chromosome:

Strand Selected:

Between DNA Start:

And DNA End:

Preproprotein Sequence Contains Sequence:

Proprotein Contains Sequence:

Signal Peptide Contains Sequence:

Cluster number:

Matches must contain a hit in the following checked parameters:

☐ salktiling ☐ blastp_plant ☐ cluster ☐ tblastn_rice

Check to display the following annotation features:

☒ chromosome
☒ strand
☒ start
☒ stop
☐ dnaseq
☐ preproprotein
☐ signalpeptide
☒ proprotein
☐ blastp_plant
☐ salktiling
☐ cluster
☐ tblastn_rice

Figure 4. Screenshot of the Arabidopsis Unannotated Secreted Peptide Database Web interface.

in our database (ath_mu_ch5_65848top [IDL2], ath_mu_ch5_8150top [IDL3], ath_mu_ch3_17023top [IDL4], and ath_mu_ch1_4155bottom [IDL5]). Finally, there are ORFs encoding peptides in this database that have high sequence similarity with annotated rice peptides (Supplemental Fig. S1).

DISCUSSION

The Arabidopsis Unannotated Secreted Peptide Database should be useful for investigators addressing a variety of questions. First, researchers studying pep-

tide biology would benefit from the list of candidate peptides for investigation. Considering that there are hundreds of orphan cell surface receptors in plants, this resource may facilitate identifying receptor-ligand pairs. Second, groups performing positional cloning of mutant genes or activation tagging would benefit from having additional gene candidates to test within a mapping interval or near the site of T-DNA insertion. Third, this resource could benefit proteomics researchers because prominent protein identification software, such as MASCOT (Perkins et al., 1999), require the protein or peptide to be in the database to make a match.

Plant Physiol. Vol. 142, 2006

general approach taken to generate this dataset have higher sensitivity to detect ORFs at the expense of lower selectivity. Based on both bioinformatic and RT-PCR data, we estimate that there are on the order of several thousand unannotated small ORFs in the human genome. Possible explanations for the ability to amplify approximately one-fifth of the tested ORFs may be that they are of extremely low abundance and would require cDNA synthesized from mRNA rather than total RNA to be detected. Alternatively, these ORFs may be induced under specific circumstances that were not tested in this study. Or, it may simply be the case that the ORFs selected to amplify are not genes.

The discrepancy in detecting ORF expression by RT-PCR and predictions of expression from tiling array data are likely due to signal-to-noise issues. If one examines tiling hybridization intensities of many annotated Arabidopsis genes, there is wide variation in the hybridization intensity within even a single exon (Yamada et al., 2003). Because the ORFs in question in this work are small, discriminating the signal above the noise is not trivial. Using tiling data to investigate small ORF expression is valuable, but the data must be interpreted with the experimental limitations kept in mind. As this resource becomes utilized by the community and further predictions are tested, estimation of the number and diversity of peptides from unannotated small genes will become clearer.

There are some limitations to our approach. Peptides that are encoded by unannotated genes with multiple exons are not represented. Also, those peptides that are produced by limited proteolysis of a

Other potential sources of extracellular peptides that are excluded from the database are those secreted by a nonclassical pathway. There is evidence in animal systems for proteins that lack signal peptides but are known to be extracellular (Bendtsen et al., 2004a). There may be a nonclassical secretory pathway operating in plants as well. For example, systemin and AtPep1 precursors apparently lack signal peptides, yet these peptides interact with cell surface receptors (McGurl et al., 1992; Scheer and Ryan, 2002; Huffaker et al., 2006; Yamaguchi et al., 2006). Second, proteomics experiments in which the cell wall or apoplast was characterized identified (in addition to proteins whose precursors do have signal peptides) many proteins that lack signal peptides (Watson et al., 2004; Boudart et al., 2005; Chivasa et al., 2005; Kwon et al., 2005). Whether the observations from proteomics work indicate impurities from the subcellular fractionation procedures used or an alternative secretory pathway is unclear (Jamet et al., 2006).

Gene identification is a prerequisite for gene investigation in the age of genomics. The Arabidopsis Unannotated Secreted Peptide Database contributes to solving the problem of small gene underannotation

and offers novel potential mediators of intercellular communication in plants that should advance our understanding of plant biology.

MATERIALS AND METHODS

Bioinformatics

To calculate protein length frequencies, TAIR 6.0 annotated proteins in FASTA format were downloaded (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR6_genome_release/TAIR6_pep_20051108). A Perl script was written to count the number of amino acids of each protein and count the frequency of protein lengths. These data were imported into Microsoft Excel and the counts of protein lengths were summed within 50-amino acid bins (e.g. 1–50, 51–100, etc.). These bin counts were plotted as a function of protein sizes represented by each bin.

For SALK tiling array data analysis, we obtained tiling array datasets (Yamada et al., 2003) from expression analyses of root, suspension cell, leaf, and flower samples (datasets kindly provided by T. Joshi and D. Xu, University of Missouri). The following steps were performed using custom Perl scripts. We removed nonrelevant features from consideration (e.g. Affymetrix controls) as well as features that were not unique (exact sequence of feature found more than once on chips). We identified the median value for each chip for the root, leaf, suspension cell, and flower experiments. We screened for features that were greater or equal to twice the median value of each chip. We then looked for an exact match of the full length of the probe feature sequence within ORF sequences. On the database, we report the probe data (raw intensity) meeting these criteria for each ORF. The feature sequence shown is the reverse complement of the actual probe sequence.

The 20-nucleotide signature Arabidopsis (*Arabidopsis thaliana*) massively parallel signature sequencing (MPSS) dataset from the Meyers' lab MPSS database (Nakano et al., 2006) was downloaded (http://mpss.udel.edu/at/public_data/20bp/20bp_summary.txt). Class 4 signatures (intergenic signatures) were obtained from the dataset and the positions of the class 4 signatures were compared with the database ORFs to identify 177 supporting signatures.

BLASTp searches were conducted using default parameters (Altschul et al., 1990) with the preproteins as queries and the plant protein database (obtained from ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/PlantProtein.Z) showing the top 10 hits, the hit scores, and the e values. tBLASTn with the predicted preproteins querying rice (*Oryza sativa* cv *japonica*), assembly version 4.0 of chromosomes by The Institute for Genomic Research (TIGR; BLAST database built from pseudochromosomes obtained from <http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml>), show best hits. Single-linkage clustering was performed using BLASTCLUST (Altschul et al., 1990) with the predicted preproteins. To aid interpretation of the database cluster data, the first number is the cluster number and the number in parentheses is the number of members in that cluster.

The ORFs and related annotation data were placed into a MySQL relational database with a searchable Web interface front end, using Perl modules DBI and CGI. The Arabidopsis Unannotated Secreted Peptide Database is publicly accessible at <http://peptidome.missouri.edu>.

RT-PCR

Total RNA was isolated from Arabidopsis ecotype Columbia plants using the Qiagen RNeasy plant mini kit according to the manufacturer's instructions (Qiagen). Two-week-old roots from plants grown on one-half-strength Murashige and Skoog agar plates under 100 $\mu\text{mol m}^{-2} \text{s}^{-1}$ yellow light were used to isolate root RNA. Leaf RNA was isolated from the rosette leaves of 3-week-old soil-grown plants. Vegetative meristem tissue was isolated from the central aerial portions of 2-week-old soil-grown plants. Flowers and fruits were obtained from 3.5-week-old soil-grown plants. Isolated RNA was treated with Turbo DNase according to the manufacturer's instructions to remove any trace amounts of genomic DNA (Ambion). RNA concentration was quantitated using a nanodrop spectrophotometer. One microgram of RNA was used to make oligo(dT)-primed first-strand cDNA with the Omniscript cDNA synthesis kit according to the manufacturer's instructions (Qiagen). Alternatively, control reactions were carried out in which the reverse transcriptase was replaced with water to the control template to rule out genomic DNA

contamination. Ten percent (2 μL) of the volume of cDNA reactions was used for 25- μL PCR reactions using gene-specific primers for 25 ORFs, which were supported by three pieces of bioinformatic data (tiling expression data belonged to a cluster and had a match to the rice genome). Primers were designed to amplify the entire ORF. TaKaRa PCR buffer (containing 1.5 mM MgCl_2 final concentration), 200 μM dNTPs, and Taq polymerase were used for PCR. PCR conditions were as follows: 95°C for 5 min followed by 36 cycles of 94°C for 30 s, 50°C for 30 s, and 72°C for 1 min. PCR products were separated on ethidium bromide containing 4% (w/v) agarose gels and photographed. The primers used for the ORFs and controls in Figure 3 were: *ath_mu_ch1_20337*top FOR: ATGGCAACTCAAGTGTCAAAGAAAATC, *ath_mu_ch1_20337*top REV: TTAAGCGGTACAATCCAGCTAAATGC, *ath_mu_ch1_22982*bottom FOR: ATGGCGCCTCAAACAATGAAAAAGAT, *ath_mu_ch1_22982*bottom REV: TTATTTTGGTTTTTACATAGCCACCAC, *ath_mu_ch3_26410*top FOR: ATGAAACAATTGTAGTCTTTCTATTG, *ath_mu_ch3_26410*top REV: TTAATAAATCCAATCATACTCTTTATGC, *ath_mu_ch4_4517*top FOR: ATGCAAACTCATAGGTGTATCGAAT, *ath_mu_ch4_4517*top REV: CTAACCGTTTATGAGGTTCTTCTTA, *ath_mu_ch4_5446*top FOR: ATGTGTGTCGCCGCATCAGGCTCACT, *ath_mu_ch4_5446*top REV: TTAATAACA-AAGGATTGTCTTAATTGAA, AT3G54350 FOR TTATCTCTTCAATTTCGA-GCCAGTG, and AT3G54350 REV: GAATCAGCAAAGAAGGATGTTTTG.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Four examples of unannotated Arabidopsis peptides aligned with their annotated rice cognates.

Supplemental Table S1. Primers used for RT-PCR of ORFs where expression was not detected.

Supplemental Table S2. Previously identified RALFL peptides in the Arabidopsis Unannotated Secreted Peptide Database.

ACKNOWLEDGMENTS

We thank Alan Marshall and Josh Hartley for help with the server, Trupti Joshi and Dong Xu for help with the tiling array data, Bill Spollen and Gordon Springer for help submitting BLAST jobs on the research cluster, the Walker lab for suggestions on the Web interface and manuscript, anonymous reviewers for their helpful feedback, and Jim Burnette for introducing Perl programming.

Received July 11, 2006; accepted September 14, 2006; published September 22, 2006.

LITERATURE CITED

- Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. *Plant Mol Biol* 60: 69–85
- Altschul SE, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410
- Basrai MA, Hieter P, Boeke JD (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res* 7: 768–771
- Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S (2004a) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel* 17: 349–356
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004b) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340: 783–795
- Boudart G, Jamet E, Rossignol M, Lafitte C, Borderies G, Jauneau A, Esquerre-Tugaye MT, Pont-Lezica R (2005) Cell wall proteins in apoplastic fluids of Arabidopsis thaliana rosettes: identification by mass spectrometry and bioinformatics. *Proteomics* 5: 212–221
- Butenko MA, Patterson SE, Grini PE, Stenvik GE, Amundsen SS, Mandal A, Aalen RB (2003) Inflorescence deficient in abscission controls floral organ abscission in Arabidopsis and identifies a novel family of putative ligands in plants. *Plant Cell* 15: 2296–2307
- Chivasa S, Simon WJ, Yu XL, Yalpani N, Slabas AR (2005) Pathogen elicitor-changes in the maize extracellular matrix proteome. *Proteomics* 18: 4894–4904

- Cobbett C, Goldsbrough P (2002) Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. *Annu Rev Plant Biol* 53: 159–182
- Cock JM, McCormick S (2001) A large family of genes that share homology with CLAVATA3. *Plant Physiol* 126: 939–942
- Fletcher JC, Brand U, Running MP, Simon R, Meyerowitz EM (1999) Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. *Science* 283: 1911–1914
- Garcia-Olmedo F, Molina A, Alamillo JM, Rodriguez-Palenzuela P (1998) Plant defense peptides. *Biopolymers* 47: 479–491
- Huffaker A, Pearce G, Ryan CA (2006) An endogenous peptide signal in Arabidopsis activates components of the innate immune response. *Proc Natl Acad Sci USA* 103: 10098–10103
- Ito Y, Nakanomoto I, Motose H, Iwamoto K, Sawa S, Dohmae N, Fukuda H (2006) Dodeca-CLE peptides as suppressors of plant stem cell differentiation. *Science* 313: 842–845
- Jamet E, Canut H, Boudart G, Pont-Lezica RF (2006) Cell wall proteins: a new insight through proteomics. *Trends Plant Sci* 11: 33–39
- Kondo T, Sawa S, Kinoshita A, Mizuno S, Kakimoto T, Fukuda H, Sakagami Y (2006) A plant peptide encoded by CLV3 identified by in situ MALDI-TOF MS analysis. *Science* 313: 845–848
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305: 567–580
- Kwon HK, Yokoyama R, Nishitani K (2005) A proteomic approach to apoplastic proteins involved in cell wall regeneration in protoplasts of Arabidopsis suspension-cultured cells. *Plant Cell Physiol* 46: 843–857
- Matsubayashi Y, Sakagami Y (1996) Phytosulfokine, sulfated peptides that induce the proliferation of single mesophyll cells of *Asparagus officinalis* L. *Proc Natl Acad Sci USA* 93: 7623–7627
- Matsubayashi Y, Takagi L, Omura N, Morita A, Sakagami Y (1999) The endogenous sulfated pentapeptide phytosulfokine- α stimulates tracheary element differentiation of isolated mesophyll cells of zinnia. *Plant Physiol* 120: 1043–1048
- McGurl B, Pearce G, Orozco-Cardenas M, Ryan CA (1992) Structure, expression, and antisense inhibition of the systemin precursor gene. *Science* 255: 1570–1573
- Moller S, Croning MD, Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17: 646–653
- Nakano M, Nobuta K, Vemuraju K, Tej SS, Skogen JW, Meyers BC (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res* 34: D731–D735
- Olsen AN, Mundy J, Skriver K (2002) Peptomics, identification of novel cationic Arabidopsis peptides with conserved sequence motifs. In *Silico Biol* 2: 441–451
- Pearce G, Moura DS, Stratmann J, Ryan CA (2001a) Production of multiple plant hormones from a single polypeptide precursor. *Nature* 411: 817–820
- Pearce G, Moura DS, Stratmann J, Ryan CA (2001b) RALF, a 5-kDa ubiquitous polypeptide in plants, arrests root growth and development. *Proc Natl Acad Sci USA* 98: 12843–12847
- Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20: 3551–3567
- Scheer JM, Ryan CA (2002) The systemin receptor SR160 from *Lycopersicon peruvianum* is a member of the LRR receptor kinase family. *Proc Natl Acad Sci USA* 99: 9585–9590
- Schopfer CR, Nasrallah ME, Nasrallah JB (1999) The male determinant of self-incompatibility in Brassica. *Science* 284: 1697–1700
- Sharma VK, Ramirez J, Fletcher JC (2003) The Arabidopsis CLV3-like (CLE) genes are expressed in diverse tissues and encode secreted proteins. *Plant Mol Biol* 51: 415–425
- Strabala TJ, O'Donnell PJ, Smit AM, Ampomah-Dwamena C, Martin EJ, Netzler N, Nieuwenhuizen NJ, Quinn BD, Foote HC, Hudson KR (2006) Gain-of-function phenotypes of many CLAVATA3/ESR genes, including four new family members, correlate with tandem variations in the conserved CLAVATA3/ESR domain. *Plant Physiol* 140: 1331–1344
- Takayama S, Shiba H, Iwano M, Shimosato H, Che FS, Kai N, Watanabe M, Suzuki G, Hinata K, Isogai A (2000) The pollen determinant of self-incompatibility in *Brassica campestris*. *Proc Natl Acad Sci USA* 97: 1920–1925
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673–4680
- Vanoosthuysen V, Miede C, Dumas C, Cock JM (2001) Two large Arabidopsis thaliana gene families are homologous to the Brassica gene superfamily that encodes pollen coat proteins and the male component of the self-incompatibility response. *Plant Mol Biol* 46: 17–34
- Watson BS, Lei Z, Dixon RA, Sumner LW (2004) Proteomics of *Medicago sativa* cell walls. *Phytochemistry* 65: 1709–1720
- Wen J, Lease KA, Walker JC (2004) DVL, a novel class of small polypeptides: overexpression alters Arabidopsis development. *Plant J* 37: 668–677
- Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, et al (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302: 842–846
- Yamaguchi Y, Pearce G, Ryan CA (2006) The cell surface leucine-rich repeat receptor for AtPep1, an endogenous peptide elicitor in Arabidopsis, is functional in transgenic tobacco cells. *Proc Natl Acad Sci USA* 103: 10104–10109